IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Speaker Recognition under Stress Conditions

Esther Rituerto-González, Ascensión Gallardo-Antolín, Carmen Peláez-Moreno

Signal Theory and Communications Department
University Carlos III Madrid

erituert@ing.uc3m.es, gallardo@tsc.uc3m.es, carmen@tsc.uc3m.es

Abstract

Speaker Recognition systems exhibit a decrease in performance when the input speech is not in optimal circumstances, for example when the user is under emotional or stress conditions. The objective of this paper is measuring the effects of stress on speech to ultimately try to mitigate its consequences on a speaker recognition task. On this paper, we develop a stress-robust speaker identification system using data selection and augmentation by means of the manipulation of the original speech utterances. An extensive experimentation has been carried out for assessing the effectiveness of the proposed techniques. First, we concluded that the best performance is always obtained when naturally stressed samples are included in the training set, and second, when these are not available, their substitution and augmentation with synthetically generated stress-like samples, improves the performance of the system.
Index Terms: speaker recognition, speaker identification, emotions, stress conditions, data augmentation, synthetic stress

1. Introduction

In recent years the interest to detect and interpret emotions in speech as well as to generate certain emotions in speech synthesis have grown in parallel. It is well-known that speech recognition systems function less efficiently when the speaker is under an emotional state, and in fact, some studies consider emotions in speech as a distortion [1].

To be able to synthesize an emotion in speech, it is necessary to analyze what are the characteristics that make it different from neutral speech. The work done about emotions in speech is very extensive, analysis are performed to study what features or combinations of them carry more information about emotions improving speech recognition rates [2], and some works aim to model emotions in speech by manipulating systematically some of the parameters of human speech, generating synthetic speech that simulates emotions [3].

Moreover, the record-keeping of databases with emotional and neutral speech is difficult as they are either recorded by actors simulating speech under those emotions, or by people under actual emotions, which could be complicated to induce. Nevertheless, stress is not considered a proper emotion, although it is intimately related to anxiety and nervousness, it is a state of mental or emotional tension resulting from adverse or demanding circumstances.

There is plenty of work about the effects of emotions in Automatic Speech Recognition (ASR) or classification of emotions in speech, but there is few work of the effects of emotions in Speaker Recognition (SR), not to mention about stressed speech on SR. Stressed speech is hard to simulate as it appears together with physical changes such as the increase of heart rate and skin perspiration. There are also hardly any databases in which stressed speech is either simulated or recorded under real conditions, along with the difficulty involved in the labelling process.

The research performed on this paper is part of a project called 'BINDI: Smart solution for Women's safety XPRIZE' by UC3M4Safety group [4]. The UC3M4Safety is a multidisciplinary team for detecting, preventing and combating violence against women from a technological point of view. The goal of this project is to develop a wearable solution that will detect a user's panic, fear and stress through physiological sensor data, speech and audio analysis and machine-learning algorithms. The ability to detect whether the voice belongs to the user or to anyone else, even under stress conditions is where this research comes in.

In this paper we want to analyze how does stress in speech affect speaker recognition rates. We aim to find techniques for strengthening speaker recognition systems, either neutralizing the effects of stress or being able to model and synthesize it from neutral speech, to create synthetically stressed speech using data augmentation techniques.

The rest of the paper is organized as follows: in Section 2 we describe the state of the art in speaker recognition and discuss features and classifiers used in literature. In Section 3, we explain the methodology followed for the feature extraction and the data augmentation techniques. Section 4 refers to the experimental set-up and results, and finally in Section 5 we discuss the conclusions and future work.

2. Speaker Recognition Related Work

Speaker Recognition is the automatic detection of a person from the characteristics of their voices (voice biometrics) [5]. We can distinguish two tasks, Speaker Identification and Verification. The first refers to the recognition of a particular user among a known number of users (a multiclass setting), and the second aims at identifying one user versus the rest (binary setting).

2.1. Features

In the literature, many features are usually used for Speaker Recognition, for example: Mel-Frequency Cepstral Coefficients (MFCC) -due to their low complexity and high performance in controlled environments-, Phonetic and Prosodic features [6] or the Linear Prediction coefficients (LP) [7]. All of these features exhibit good performance in the task when used in neutral or emotionless speech.

For speaker recognition under stress conditions, however, there is hardly any previous work, even though, MFCCs, along with Linear Frequency Cepstral Coefficients (LFCC) and Linear Prediction Cepstral Coefficients (LPCC) are cited as important features [8], together with the Pitch, Energy and Duration, which are features that seem to differ between speakers.

10.21437/IberSPEECH.2018-4